New Frontiers in Business Intelligence

Surajit Chaudhuri
Microsoft Research
One Microsoft Way
Redmond, WA, 98052
surajitc@microsoft.com

Vivek Narasayya
Microsoft Research
One Microsoft Way
Redmond, WA, 98052
viveknar@microsoft.com

1. INTRODUCTION

Business intelligence (BI) software is a collection of decision support technologies for the enterprise aimed at enabling knowledge workers such as executives, managers and analysts to make better and faster decisions. The past two decades have seen explosive growth, both in the number of products and services offered and in the adoption of these technologies by industry. This growth has been fueled by the declining cost of acquiring and storing very large amounts of data arising from sources such as customer transactions in banking, retail as well as in E-businesses, RFID tags for inventory tracking, email, query logs for websites, blogs and product reviews. Enterprises today collect data at a finer granularity which is therefore of much larger volume. Businesses are leveraging their data assets aggressively by deploying and experimenting with more sophisticated data analysis techniques to drive business decisions and deliver new functionality such as personalized offers and services to customers. Today, it is difficult to find a successful enterprise that has not leveraged BI technology for their business. For example, BI technology is used in manufacturing for order shipment and customer support, in retail for user profiling to target grocery coupons during checkout, in financial services for claims analysis and fraud detection, in transportation for fleet management, in telecommunications for identifying reasons for customer churn, in utilities for power usage analysis, and in E-business for identifying customers who are likely to respond to a product catalog mailing campaign.

When compared to the BI landscape of the mid-90s (see [1] for a survey from that period), we observe that today's BI technology has progressed well beyond OLAP servers, parallel DBMSs, and classical ETL of that era. Indeed, several new frontiers in BI have emerged: (a) "*Big Data*" engines. Two new data parallel architectures are gaining popularity for enterprise analytics. First, *data warehouse appliances* offer the promise of reduced cost through a combination of packaged hardware and software that is pre-installed and pre-configured for data warehousing. Second, BI platforms based on the *MapReduce* paradigm are also beginning to gain traction for analytics at large scale. (b) The need to shorten the time between data acquisition and making business decisions has led to developments in *near real-time BI* technology. (c) *Predictive analytics* on structured as well as text data has gained in importance as enterprises today are even more willing to consider deeper analysis techniques. (d) *Enterprise search* engines have emerged as a gateway to and to help unlock value from diverse sources of data (e.g. databases, email servers, text documents) available within an enterprise. (e) *Cloud data services* are beginning to provide hosted BI services that are aimed at reducing manageability burdens such as provisioning and availability.

While most of the recent innovation in BI technology has been driven by the industry, the academic community has an important role to play as many important and non-trivial research challenges remain. This presentation describes scenarios and key enabling technologies for BI, outlines new challenges and opportunities, and reflects on recent hardware and application trends influencing BI technology. We will first provide a broad overview of the current BI landscape identifying important use cases, describing key technologies, and highlighting relationships across these technologies. We expect that the broad overview of BI will be informative for researchers and practitioners alike. Next, we drill-down into five newly emerging frontiers of BI discussed above, where there has been large growth recently and relatively little emphasis in the research community. For each area we discuss state-of-the-art, highlight architectural considerations and discuss open research problems. For this tutorial, we will leverage material from our upcoming article in Communications of the ACM [2]. The rest of this document summarizes the key elements of our presentation.

2. OVERVIEW OF BI TECHNOLOGY

Our overview of the BI technology will illustrate the breadth of BI technologies (see Figure 1). Important examples of BI technology include techniques for *data movement* from operational databases into data warehouses including ETL (e.g. data cleansing, data loading) and Complex Event Processing engines that analyze streaming data before it enters the data warehouse.

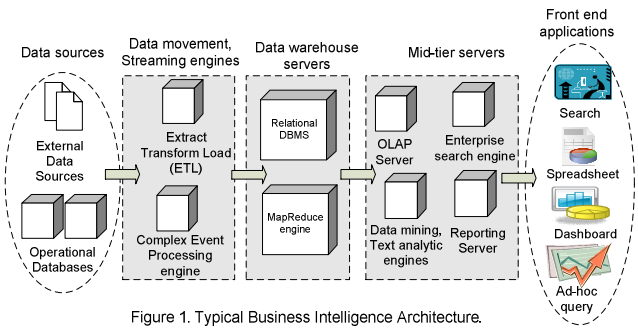

Figure 1. Typical Business Intelligence Architecture.

One other key development has been changes in the landscape of *data warehouse server* technology from traditional parallel DBMS technology that include progress in topics such as data

Proceedings of the VLDB Endowment, Vol. 4, No. 12

compression and column-oriented storage. Another important facet of BI technology is functionality and architectures for mid-tier servers that support specialized analytics: OLAP servers, enterprise search engines, data mining, text analytic engines and reporting servers. We also discuss front end tools and applications that are widely used by decision makers within the enterprise.

3. "BIG DATA" ENGINES

Enterprises are analyzing large amounts of data (e.g. sensor data, text documents from a web crawl, log and clickstream data) in order to realize competitive differentiation. In recent years, two alternative architectures have emerged to support such "big data" analytics: *data warehouse appliances* and *MapReduce engines*. In brief, a data warehouse appliance is an integrated set of storage hardware, operating system and DBMS software specifically *pre-installed* and *pre-optimized* for data warehousing. Appliances can also push part of the query processing into specialized hardware thereby speeding up analytic queries significantly. The motivation for the *MapReduce* paradigm originally came from the need to have a scalable infrastructure for index generation and analytics over query logs and web data. More recently however, such engines are being extended to support a more general class of data parallel execution of tasks, thereby targeting traditional data warehousing scenarios.

4. NEAR REAL-TIME BI

The competitive pressure of today's businesses has led to the increased need for *near real-time* BI. The goal of near real-time BI (also called *operational* BI or *just-in-time* BI) is to reduce the latency between when operational data is acquired and when analysis over that data is possible. Consider an airline that tracks its most profitable customers. If a high-value customer has a lengthy delay for a flight, alerting the ground staff proactively can help the airline ensure that the customer is given priority for re-routing. Such near real-time decisions can increase customer satisfaction and revenue. There are two alternative approaches for supporting near real-time BI: *Complex Event Processing (CEP) Engines* and techniques for *fast ETL*. CEP is different from traditional BI since operational data does not need to be first loaded into a warehouse before it can be analyzed. Businesses can specify the patterns or temporal trends that they wish to detect over streaming operational data (referred to as events), and take appropriate actions when those patterns occur. Another approach for enabling near real-time BI is to make the ETL process very fast by using efficient techniques for capturing changed data and loading.

5. ENTERPRISE SEARCH

BI tasks often require searching over different types of data within the enterprise. For example, a salesperson who is preparing for a meeting with a customer would like to know relevant customer information before the meeting. This information is today siloed into different sources: e.g. CRM databases, emails, documents, and spreadsheets, both in enterprise servers as well as on the user's desktop. The ability to retrieve and rank the required information using the popular keyword search paradigm is valuable for BI. Enterprise search focuses on supporting the keyword search paradigm over text repositories and structured enterprise data. Today, a number of vendors provide enterprise search capability. We will present the architecture used in the popular integrated model used in today's enterprise search engines. We also discuss technical challenges in crawling data sources, indexing and query processing components of enterprise search engines.

6. PREDICTIVE ANALYTICS

Predictive analytics plays an increasingly important role in enterprise BI and can help answer deep analytic questions. For example, an e-tailer might want to know: (a) Which of my customers are likely to respond to my upcoming product catalog mailing campaign? (b) Based on the text of a recent survey of products sold by the company, do customers view the product positively or negatively? We will discuss use cases, technologies, and open problems in using machine learning and other analytic techniques over structured as well as text data. Our discussions will include techniques for extracting structured data from text documents, that can be broadly categorized as: (1) *Named entities* such as locations, people, products, organizations etc. (2) *Concepts/topics.* (3) *Sentiment analysis* (e.g. label each document with a "positive", "neutral" or "negative" sentiment).

6.1 CLOUD DATA SERVICES

Managing enterprise BI today requires handling tasks such as hardware provisioning, availability, and security patching for servers used to support BI. The success of hardware virtualization in the cloud has prompted database vendors to virtualize *data services* so as to further improve resource utilization and reduce cost. These data services initially started as simple key-value stores but have now begun to support the functionality of a single node relational database as a hosted service. While the intended initial users of such cloud database services are relatively simple departmental OLTP applications, there is potential to extend the paradigm to BI as well. In our presentation, we will touch upon cloud data service architectures and the unique challenges that arise for them for supporting the full range of BI services.

7. PRESENTERS

Surajit Chaudhuri is a Research Manager at Microsoft Research, Redmond. He started the AutoAdmin project on self-tuning database systems at Microsoft Research. Surajit has also worked in the area of data cleaning. His research on both physical database design and data cleaning has been incorporated in Microsoft products and services such as Microsoft SQL Server and Bing. Surajit received his Ph.D. from Stanford University and is an ACM Fellow. He was awarded the ACM SIGMOD Contributions award in 2004, a VLDB 10-year Best paper Award in 2007 and the ACM SIGMOD Innovations Award in 2011.

Vivek Narasayya is a Principal Researcher at Microsoft Research, Redmond. He is interested broadly in data management, focusing on the areas of self-tuning database systems, query processing, query optimization, and resource management in databases. He did his Ph.D. from the University of Washington, Seattle. He was awarded a VLDB 10-year Best paper Award in 2007.

8. REFERENCES

[1] Chaudhuri S. and Dayal U. *An Overview of Data Warehousing and OLAP Technology*. SIGMOD Record, 26(1) pp. 65-74, 1997.

[2] Chaudhuri S., Dayal U., Narasayya V. An Overview of Business Intelligence Technology. To appear in Communications of the ACM. (CACM), 2011.